Objects and scenes classification with selective use of central and peripheral image content

Objects and Scenes Classification with Selective Use of Central and Peripheral Image Content

Ali Alameer[a,**], Patrick Degenaar[a,b], Kianoush Nazarpour[a,b,**]

[a] *School of Engineering, Newcastle University, Newcastle NE1 7RU, UK*
[b] *Institute of Neuroscience, Newcastle University, Newcastle NE2 4HH, UK*

Abstract

The human visual recognition system is more efficient than any current robotic vision setting. One reason for this superiority is that humans utilize different fields of vision, depending on the recognition task. For instance, experiments on human subjects show that the peripheral vision is more useful than the central vision in recognizing scenes. We tested our recently-developed model, that is, the elastic net-regularized hierarchical MAX (En-HMAX), in recognizing objects and scenes. In various experimental conditions, images were occluded with windows and scotomas of varying sizes. With this model, classification accuracies of up to 90% for objects and scenes were possible. Modelling human experiments, window and scotoma analysis with the En-HMAX model revealed that object and scene recognition are sensitive to the availability of data in the centre and the periphery of the images, respectively. Similarly, results of deep learning models have shown that the classification accuracy diminishes dramatically in the absence of the peripheral vision. These differences led us to further analyse the performance of the En-HMAX model with the parafoveal versus peripheral areas of vision, in a second study. Results of the second study show that approximately 50% of the visual field would be sufficient to achieve 96% accuracy in the classification of unseen images. The En-HMAX model adopts a relative order of importance, similar to the human visual system, depending on the image category. We showed that utilizing the relevant regions of vision can significantly reduce the image processing time and size.

Keywords:
Visual recognition
Image understanding
Biological visual-systems
Visual perception
Scene analysis

*Corresponding author: Tel.: +44 (0) 191 208 3337
**Corresponding author: Tel.: +44 (0) 191 208 6860
Email addresses: ali.alameer@newcastle.ac.uk (Ali Alameer), kianoush.nazarpour@newcastle.ac.uk (Kianoush Nazarpour)

1. Introduction

Humans can categorize objects and scenes within 100 ms, despite variations in pose, size, and lighting conditions [75, 50]. This *feedforward* process is partially underpinned by the hierarchical structure of the visual cortex [68]. Recent data-driven computer approximations of human vision [66, 42, 22, 82] have attempted to achieve this level of performance with varying success. Understanding the neural mechanisms that underlie the categorization of objects and scenes may pave the way for the development of robust machine vision systems.

One key feature of the human visual system is that it processes the peripheral and central information of the visual field in parallel [30] via specialized structures within the visual cortex [32]. For example, functional brain imaging showed more brain activity in fusiform face area (FFA) when categorizing data, which appear more likely at the center of the visual field, such as faces [39] and words [58]. However, more activity was registered in the parahippocampal place area (PPA) during recognition of scenes, such as buildings, which lie more likely at the periphery of the visual field [21, 61]. It has been suggested that the mid-fusiform sulcus (MFS) area of the brain enables this fast parallel processing by segregating the peripherally- and centrally-biased pathways [30].

Another very interesting feature of the human visual system is that it strikes a trade-off between different fields of vision and their resolution via a process called foveation [69, 57, 15]. Specifically, it reduces the *neural* processing resolution in the peripheral vision [12]. As a consequence, the highest level of object recognition accuracy can be achieved within the visual angles[1] of $[1^\circ - 2^\circ]$ of the fixation point [34, 41]. Whilst the accuracy drops gradually as the object moves away from the fixation point, because of this neural compression, the speed of visual processing improves [26, 14]. In addition, this structural setting allows the peripheral vision to be more sensitive in recognizing scenes [11]. Foveation was integrated with convolutional neural networks (CNNs) to mitigate the effect of the adversarial perturbations [55]. However, the impact of foveation on other biologically-inspired models on human vision is understudied.

Recent advances in object detection involved introducing a simple method to train rotation-invariant and Fisher discriminative CNN models [16] to boost CNN performance. Another study proposed a rotation-invariant CNN (RICNN) model using a new rotation-invariant layer embedded within the architecture of an existing CNN [18]. In order to enhance the performance of existing methods of remote sensing image scene classification, a recent study proposed a simple approach to learn discriminative CNNs (D-CNNs) [17]. A multitask model that merge scene images of different resolutions was proposed [53]. The model selects the optimal information and preserves the underlying manifold structure of data by using a sparse feature selection-based manifold regularization (SFSMR). Moreover, an unsupervised representation learning method was proposed to investigate deconvolution networks for remote sensing scene classification [54]. A

[1] By definition, the angle formed by the two extremities of a viewed object or scene is referred to as the visual angle [38].

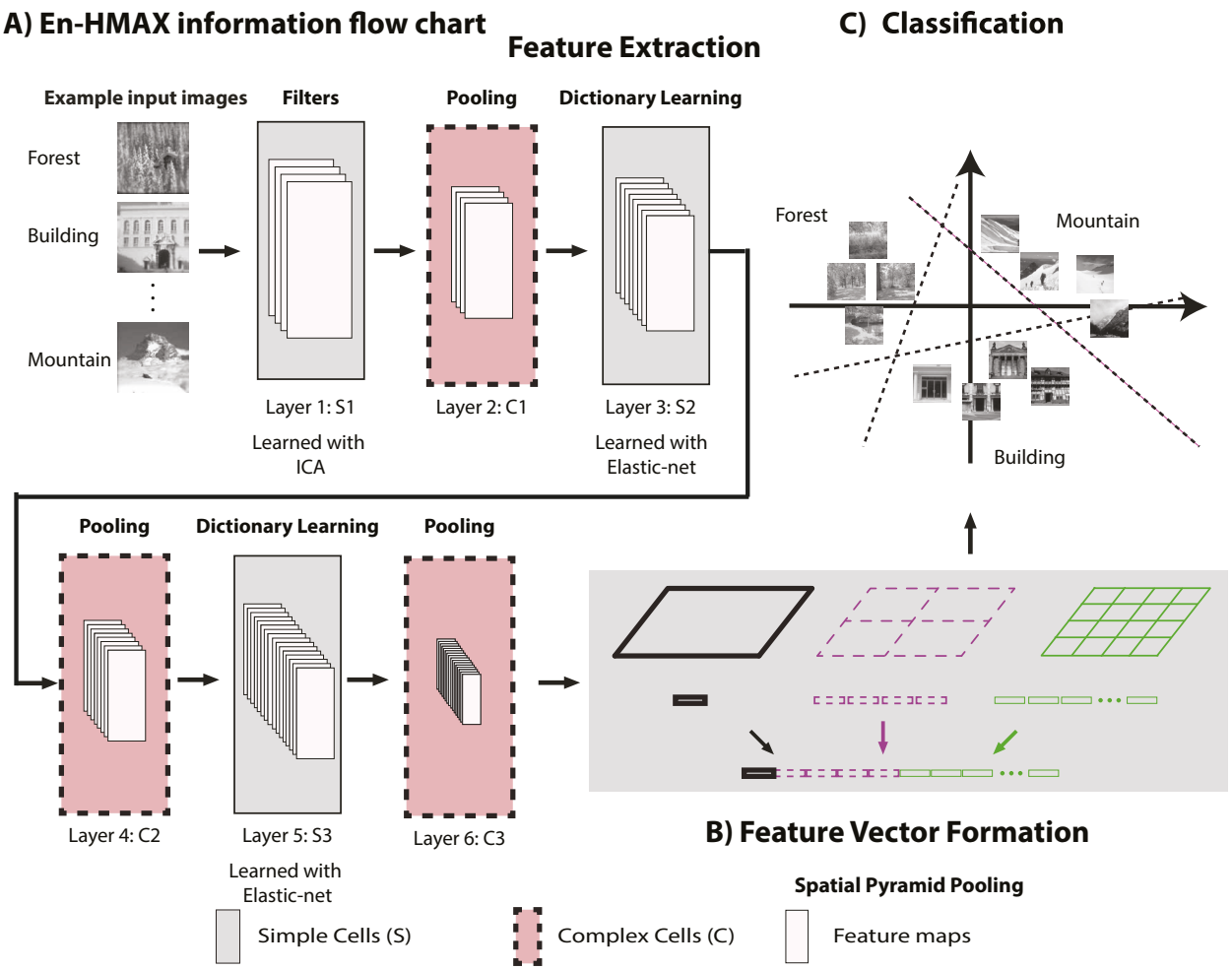

Figure 1: A) Schematic of the En-HMAX model with each block representing an S or C layer of the model along with their function. B) Spatial pyramid pooling layer with a grid resolution of {1, 2, 4}. C) The classification layer that include a one-versus-all linear SVM classifier.

shallow weighted deconvolution network was used to learn a set of filters and feature maps for each image. Experimental results show that this approach outperforms most state of the arts. Similarly, to further enhance remote sensing scene classification, a bidirectional adaptive feature fusion strategy was developed [52]. The SIFT feature is fused with the deep learning feature to produce discriminative image presentation.

Wang and Cottrell [78] investigated the advantages of peripheral vision in scene recognition. In their experiments, they used deep learning methods to replicate the experimental environment of Larson and Loschky [45]. They showed that the peripheral advantage emerges naturally in the learning process. When trained to categorize scenes, the model weights the peripheral pathway more than the central pathway.

Computational models of the mammalian visual cortex were intitially inspired by the early work of Hubel and Wiesel [37]. A hallmark of their research is that neurons of the visual cortex extract successively

complex information from the pattern of observed objects. More recent studies showed that the point spread function of neurons in primates can be modelled using Gaussian derivative filters with different scales and orientations [81]. In addition, more accurate filters can be modelled by learning statistics from natural scene images [64]. Cortex-inspired models, such as the hierarchical MAX (HMAX) model [68], simulate the ventral pathway of the visual cortex [10] with a hierarchy of simple and complex cells. In the HMAX model, the input data is sparsified through an alternate convolutional and pooling layers, allowing selectivity and invariance for preserving the patterns of objects [68]. Similarly, the En-HMAX model was developed to mimic basic structures of the ventral visual system; a hierarchy of brain areas mediate object recognition. It model the first 100 ms of the feedforward visual cognition of primates [7]. It differs from deep learning methods in that it is feed-forward, in terms of training/processing data, with no back-propagation or feedback loops. Studies suggest that the correspondences of convolutional neural network (CNN) to the structures and mechanisms of the visual cortex are not quite clear [49].

In this paper, we test our model of human vision, that is the elastic net-regularized hierarchical MAX (En-HMAX) for two specific features of human vision:

- Flexible utilization of peripheral versus central vision to enhance scene and object recognition performance.
- Central foveation to reduce the size of the visual data without compromising the recognition performance.

Computationally, this paper investigates the trade-off between the processing time, image-size, and accuracy when utilizing effective regions of vision. The main novelties of this work are:

1. Modelling human experiments with utilising the En-HMAX model to quantify the effectiveness of peripheral and central vision by applying foveation, scotoma and window conditions.
2. Investigating the behaviour of recent deep learning methods using the above experimental environment.
3. Analysing the trade-off between the classification accuracy and computational requirements, i.e., time and data size, to process an image given its category

2. Methods

2.1. The Hierarchical MAX (HMAX) model

The HMAX model is a computational model that summarizes the basic facts about the ventral visual stream of the primate's visual cortex. It comprises layers of simple (S) and complex (C) cells which are configured into four primary layers, namely: S_1, C_1, S_2 and C_2. The S_1 unit is a set of Gabor filters $F(x, y)$, resembling the receptive fields in cortical simple cells. The S_1 layer preprocesses the input image.

The complex C_1 units obtain the maxima of neighboring square patches $\mathbf{u}_{i,j}$ of S_1 feature maps to increases the invariance to transformations and translation. C_1 patches are compared with prototypes such that smaller distances elicit larger responses. Finally, the C_2 layer response is generated by "max-pooling" of S_2 response for position- and scale-invariant feature maps [24].

2.2. The En-HMAX Model

Our En-HMAX model (Figure 1A) [8, 7, 4, 5, 9] comprises three layers, each consisting of both simple S and complex C units. We use independent component analysis (ICA) to generate filters that resemble the receptive fields of V1 simple cells. Extracting filters from natural images using ICA is believed to better model V1 receptive fields of the visual cortex [40]. The S_2 and S_3 units of the En-HMAX model feature an elastic-net regularized dictionary learning [85, 1] to reinforce model sparsity and grouping effect, simultaneously. Let $\mathbf{x_i} \in \Re^m$ be an image patch extracted randomly from C_1 or C_2 units and introduced to the S_2 or S_3 units,respectively, where m denotes the size of the image patch. Given a set of bases $\mathbf{d_i} \in \Re^m$, sparse coding searches for the sparse coefficients s_j such that $\mathbf{x_i} = \sum_{j=1}^{p} \mathbf{d_i}\ s_j$, where p denotes the size of the dictionary. Therefore, we have $\mathbf{X} = \mathbf{DS}$, where $\mathbf{X}$ is an n-dimensional local descriptor extracted from the input images and $\mathbf{D}$ is a p-dimensional dictionary matrix, where n corresponds to the selected number of image patches to be extracted from the previous layer. Each column of $\mathbf{S}$ is a vector $\mathbf{s_i} \in \Re^p$ holding the sparse coefficients. As such, the elastic-net formulation is

$$\begin{aligned} \text{minimize} \quad & \|\mathbf{X} - \mathbf{DS}\|_F^2 + \lambda_1 \|\mathbf{S}\|_1 + \lambda_2 \|\mathbf{S}\|_F^2 \\ \text{subject to} \quad & \|\mathbf{d}_i\|_2 \leq 1, \forall i = 1, ..., p, \end{aligned} \tag{1}$$

where $\|.\|_F$, $\|.\|_2$ and $\|.\|_1$ denote the Frobenius, ℓ_2 and ℓ_1 norms, respectively. The penalty functions encourage the model to introduce sparsity and grouping effect when processing highly correlated image data. For every input image patch $\mathbf{x_i}$ in $\mathbf{X} \in \Re^{m \times n}$, a vector $\mathbf{s_i}$ in $\mathbf{S} \in \Re^{p \times n}$ is reproduced, corresponding to a basis $\mathbf{d_i}$ in the dictionary $\mathbf{D} \in \Re^{m \times p}$. The sparsity of the coefficients is controlled by λ_1. The scalar λ_2 controls the sensitivity of basis selection from the dictionary. Following [85], cross-validation was used to tune λ_1 and λ_2.

The C layers perform ℓ_1 norm-pooling, that is, higher-level units are assigned selective responses from the preceding lower-level units to enhance invariance against translation and scaling [33]. The En-HAMX structure is followed by a feature formation layer in which the spatial pyramid pooling (SPP) technique [79] is used. With the SPP method, with a grid resolution of {1, 2, 4}, each feature map in C_3 is transformed into a feature vector of length 21, (Figure 1B). A classifier constitutes the final stage (Figure 1C).

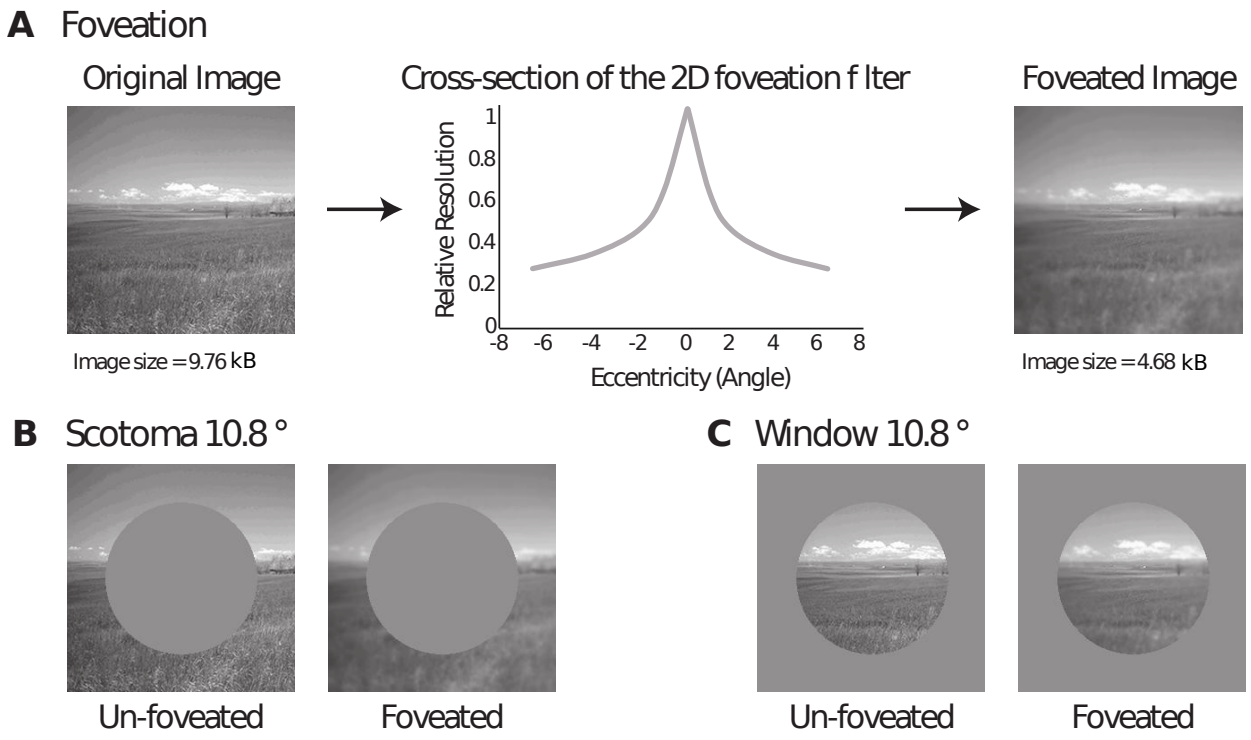

Figure 2: An example of pre-processing an image [from [48]] with Foveation, scotoma and window conditions. (A) Foveating an image using a 2D filter; (B) Examples of the scotoma condition; (C) Examples of the window condition.

2.3. Experiments

2.4. Scenes and Objects Image Datasets

To test the En-HMAX model, we utilized both scene and object image datasets. The scenes dataset included human-made and natural scenes. The images of the scene dataset were collected from recent scene image datasets [48, 46, 84]. The classes of our objects database were extracted from the Caltech 101 database [23] and ImageNet dataset [20]. Overall, the scenes and object datasets included 15 and 11 classes, respectively; with a total of 14 million images to pre-train, train and test our models (see Appendix). The images of the dataset were 300 $\times$ 250 pixels, on average.

2.5. Images with scotoma and window

By definition [59], the term *window* is originated by the analogy of looking at a scene through a window. In the window paradigm, the visual information outside the window is absent. The term *scotoma* is derived from an analogous medical condition in which a certain part of the visual field is blocked. In the scotoma paradigm, the information outside the scotoma is unaltered and the centre-based information is blocked. The window and scotoma paradigms have been utilized regularly in understanding the mechanisms underpinning human vision, for example, [59, 35, 25, 76, 65]. In addition, the scotoma and window were utilized in the

experiments of [45] to demonstrate recognition accuracy to investigate the contribution of peripheral versus central vision. Therefore, a scotoma and a window were used to investigate the performance of the En-HMAX model in the classification of occluded images. Following [35], they were constructed with:

$$h_g(n,m) = \exp\left(\frac{-(n^2+m^2)}{2\sigma^2}\right) \tag{2}$$

$$h(n,m) = \frac{h_g(n,m)}{\sum_n \sum_m h_g} \tag{3}$$

where σ denotes the standard deviation which acts as a threshold that determines the boundaries of the mask, $h_g(\cdot,\cdot)$ corresponds to the distribution function, $h(\cdot,\cdot)$ is the generated normalized multivariate Gaussian kernels, and (n,m) represent the dimension of the kernel. We then discretized the mask by setting all pixel values inside (or outside) the mask to 128 to form the scotoma (or window). Examples for applying window and scotoma on an image in its original and foveated forms are shown in Figure 2(B, C). In Experiment 1, in line with [45], we set the distance between the model and the images to 70 cm. In Experiment 2, we used a wider range for σ.

2.6. Foveation

We measured the effect of foveation on the performance of the En-HMAX model. We used a pyramid of low pass filters [27]. Each input image, e.g. Figure 2A, was passed through six repeated layers of filters cascaded with a down-sampling stage. Starting from the centre of each image, we set the filtering and down-sampling parameters such that at each pyramid layer the image resolution was halved [27]. We applied foveation to input images such that the contrast c is calculated with: $c(f,e) = c_0 \exp(\alpha f \frac{e+e_2}{e_2}$ where f is the spatial frequency, c_0 is the minimum contrast threshold, e is the retinal eccentricity, e_2 is the half-resolution eccentricity and α is the decay constant. Figure 2A illustrates foveation of an image.

2.6.1. Experiment 1

We quantified the performance of the En-HMAX model in classification of various scenes and objects data under window and scotoma occlusion conditions. This experiment comprised two parts as shown in Figure 3A. In part one, we classified the original images of the scene and object datasets. In part two, all images were first foveated before repeating the analysis exactly as in part 1. In both parts, we trained the En-HMAX model with original images and tested it with a fixed number of images; but overlaid the images with windows or scotomas of four visual angles: 1°, 5°, 10.8° and 13.6°. These angles for the window and scotoma modelled [45] the presence and absence of foveal information (1°), parafoveal and foveal vision against the peripheral vision (5°), and peripheral information (13.6°). The case of 10.8° represented the situation of equal areas inside the window and outside the scotoma.

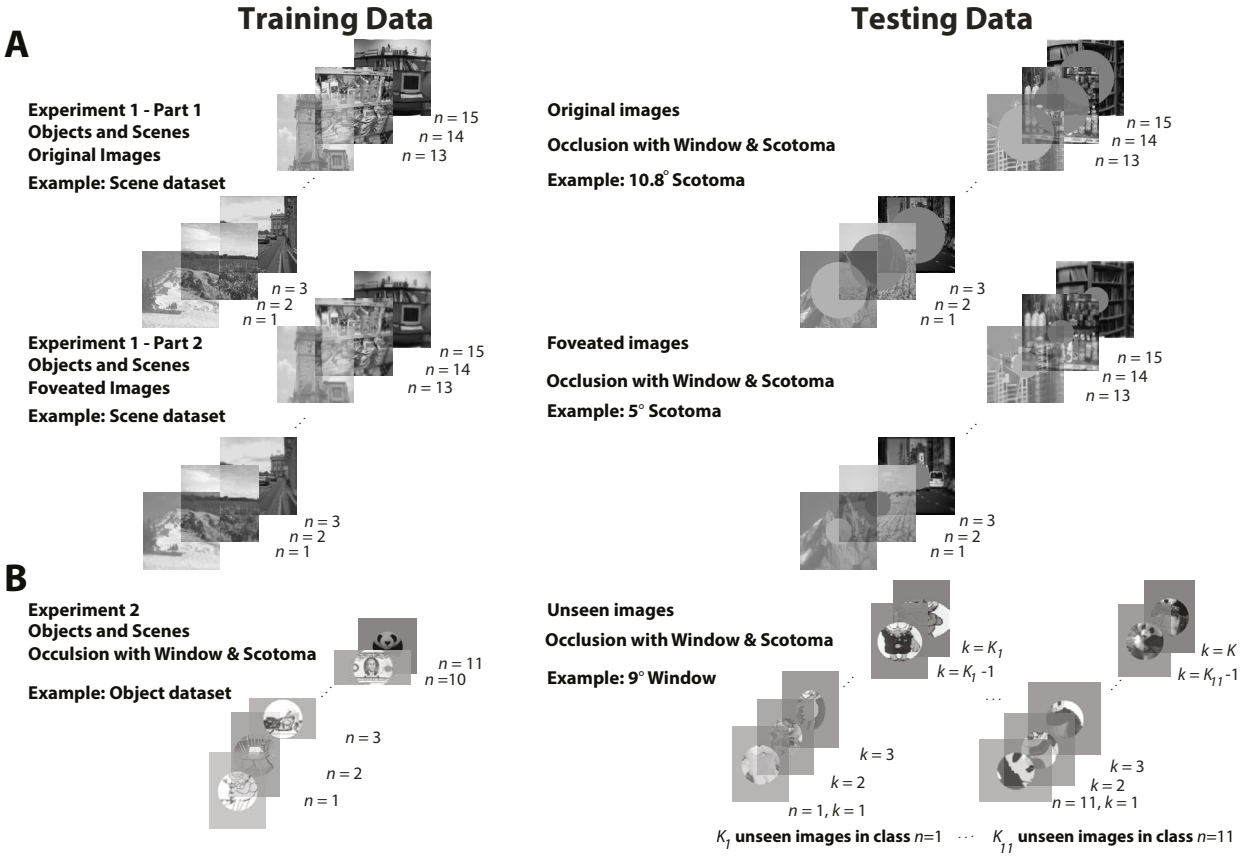

Figure 3: The configuration of the experiments. Similar settings have been used for Experiment 1 and Experiment 2. In Experiment 2, the number of testing images varies, depending on the size of each class. The letters n and k represent the class number and the image number in the each class, respectively.

2.6.2. Experiment 2

We tested the capability of the En-HMAX model, trained with occluded images, to generalize to unseen images. The configuration of this experiment allows identifying, on a micro-level, the effective regions of vision as a result of training and testing the model with similar occlusion type (scotma or window) and visual angles. We measured the efficiency of each region of vision in both datasets at visual angles $[1^\circ - 19^\circ]$ with step size 2°, and the model was tested with unseen images for all occlusion angles. An example of this classification design for the objects dataset is shown in Figure 3B.

2.7. Classification

For the En-HMAX model, we used the one-versus-all method for multi-class linear support vector machine (SVM) classification. On the other hand, we utilized the output of the softmax layer for classification in the case of CNN models.

For CNN models, we used networks that were pre-trained with scene images, Places dataset [84], and

object images, ImageNet dataset [20] depending on the classification task. We adjusted the CNN models to our dataset configuration by replacing the last fully convolutional layer. We froze the weights of the first ten layers and only retrained the weights of the advanced layers. We used Adam optimizer [43] without applying any image augmentation to the input data.

In Experiment 1, we randomly selected 100 images per category to train the En-HMAX model; thirty images per category were chosen randomly to measure the overall error rates. We repeated this process 20 times. We used an ample number of images to train and test the models to allow for a precise investigation for the contribution of each region of vision. Furthermore, we did not introduce any image augmentation to increase the training data. We obtained the performance of three convolutional neural networks, namely, AlexNet [44], VGG19 [70], GoogLeNet [72], and MobileNet-v2 [67] along the En-HMAX model.

In Experiment 2, we used 10-folds cross-validation to train and test the En-HMAX model. We included all the images of both datasets (see Appendix). Finally, we calculated the average accuracy and standard deviation across all ten folds.

3. Results

3.1. Experiments 1

Figure 4 shows the results of Experiment 1 (both parts); in which the En-HMAX model was trained with complete images and was tested on images with window or scotoma occlusions.

3.1.1. Scene classification

Figure 4A reports the results of scene classification. In the cases of 1° and 5° scotoma, accuracies of approximately $89 \pm 1\%$ were achieved. The difference in the scores for 1° scotoma ($M = 89.2, SD = 5.6$) and 5° scotoma ($M = 88.9, SD = 4.2$) was not significant: (non-parametric sign test, $z_{19} = 0.8, p = 0.3$), where M and SD denote the mean and standard deviation, receptively. This finding indicates that the En-HMAX model can achieve the maximum performance even in the absence of parafoveal vision (5°). On the other hand, the accuracy at 13.6° scotoma reduced to $23 \pm 6\%$. The performance was poor at 1° and 5° visual angle window conditions. This score increased as the window expanded. At the 13.6° window condition, the accuracy reached 57%.

In the 10.8° scotoma condition, the difference in the scores for original ($M = 37.1, SD = 12$) and foveated images ($M = 27.4, SD = 11.1$) was significant; (paired samples t-test, $t_{19} = 3.0, p = 6.5 \times 10^{-3}$). Similarly, for 13.6° scotoma condition, the difference in the scores for original ($M = 21.6, SD = 7.9$) and foveated images ($M = 15.9, SD = 4.8$) was significant; (paired samples t-test, $t_{19} = 2.6, p = 0.016$). In the window condition, there was no significant difference in classification of the original and foveated images. Two examples for scene images are presented in Figure 4C. Foveation has significantly reduced the resolution of the scene outside the scotoma and led to degradation of performance.

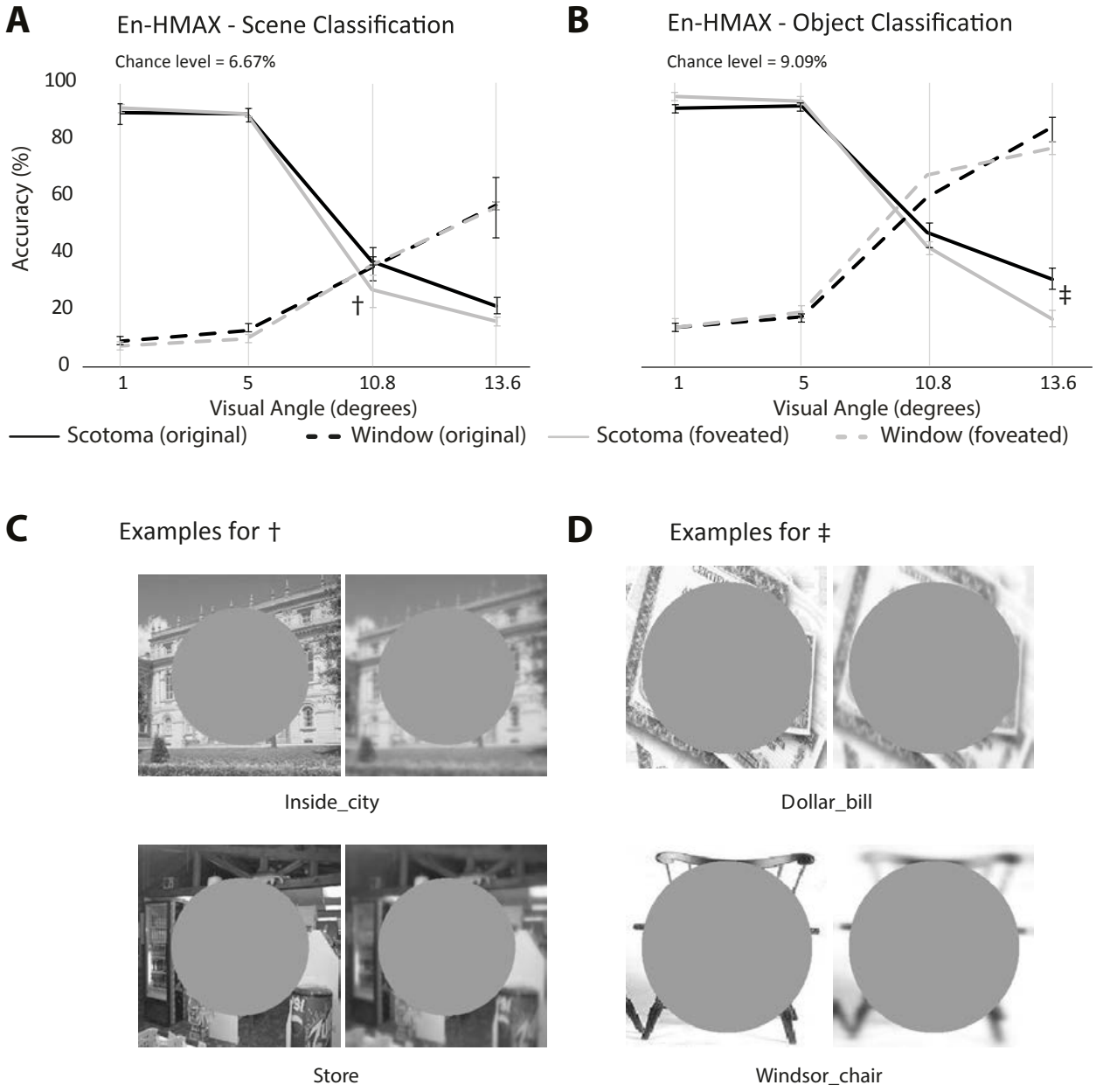

Figure 4: Classification accuracy with the En-HMAX model as a function of visual angle and viewing condition (scotoma and window) for scene (A) and object (B) images with and without foveation. (C), (D) Examples for the 10.8° and for 13.6° scotoma condition for both original and foveated images.

3.1.2. Object Classification

In the object classification, Figure 4B, classification accuracies exhibited similar trends to Figure 4A. However in comparison, several interesting features were observed.

For objects, the cross-over of window and scotoma conditions occured at visual angles of 9.7° (original)

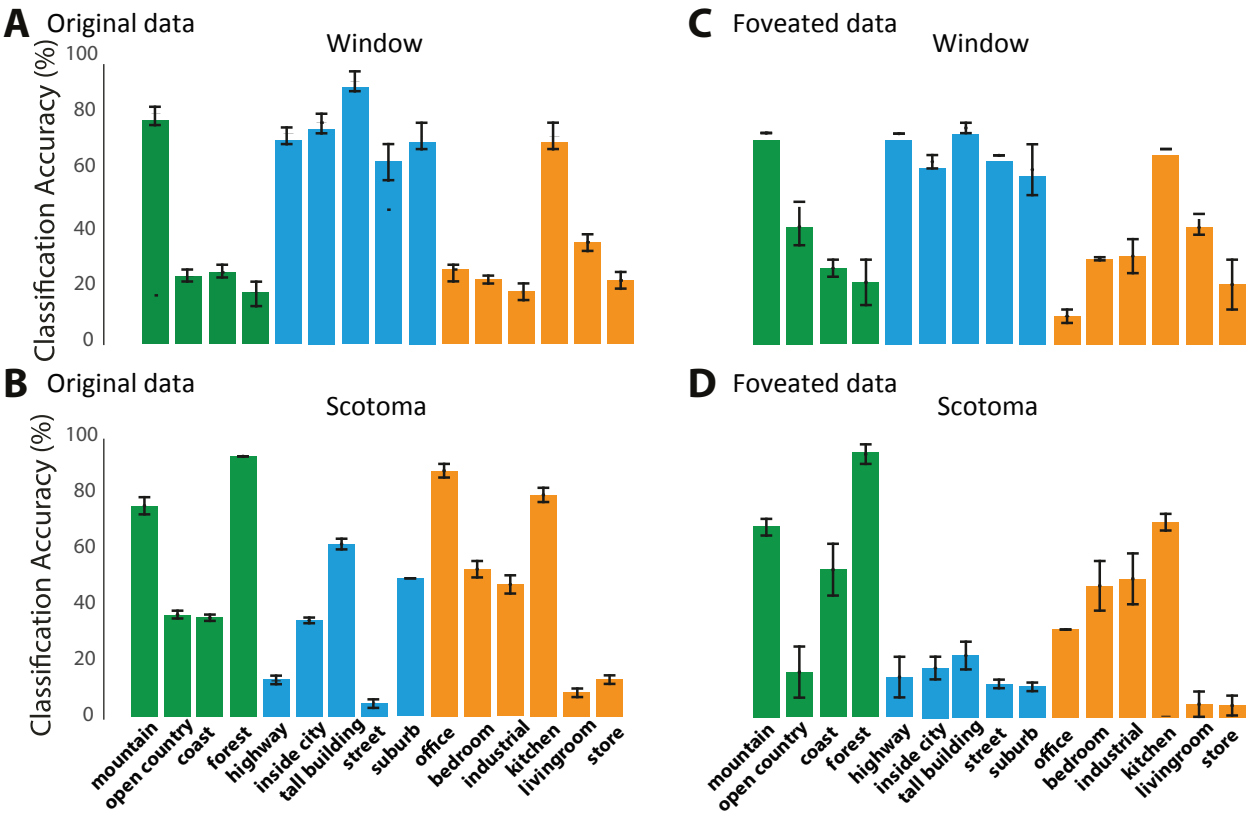

Figure 5: Individual class accuracies for the scene dataset at an angle of 10.8° in the window and scotoma conditions. The classes are categorized according to whether they are natural (green), man-made and out-door (blue) or man-made and in-door (amber) scenes.

and 9° (foveated). However, for scenes, it was at 10.8° (original) and 10° (foveated) (Figure 4A). This observation indicates that the En-HMAX model relies more on the central image content for recognising objects.

At the 13.6° window condition, the classification performance of objects ($M = 84.7, SD = 7.6$) was significantly higher than that of scenes ($M = 57.1, SD = 25.5$); (non-parametric sign test: $z_{19} = 2.6, p = 7.2 \times 10^{-3}$). This indicates that the peripheral region of the scene images is more effective for the recognition process.

At the 1° scotoma condition, the objects classification score achieved for the foveated images ($M = 95.1, SD = 3$) was significantly higher than that observed for the original object images ($M = 91.2, SD = 2.1$), (paired samples t-test, $t_{19} = 2.7, p = 0.01$). This result confirms that the En-HMAX can benefit from foveation in categorizing objects because the relevant spatial content is dense near the object and sparse in the surrounding vicinity. However, at the 13.6° scotoma condition, the objects classification score achieved for the foveated images ($M = 17, SD = 6.2$) was significantly lower than that observed for the

original images ($M = 31.1, SD = 6$); (paired samples t-test, $t_{19} = 6.9, p = 1.32 \times 10^{-6}$). This observation was unexpected as it was believed that foveation should not affect object categorization in the scotoma condition. We speculate a reason for this observation can be that some objects occupy the whole image, e.g. the two examples presented in Figure 4D. As such, the object information that fall outside the scotoma, may still be useful in classification. When they are blurred by foveation, the classification score deteriorates.

3.1.3. Scene subtypes

Figure 5 shows the accuracies of individual classes in the scene dataset. We categorized the scenes according to whether they were natural (green), man-made and out-door (blue) or man-made and in-door (amber) scenes. We show only the following viewing conditions to both the original and the foveated version of the images: window 10.8° and scotoma 10.8°. The rationale of selecting only these two conditions was to observe how scene classification was affected when central or peripheral image content was blocked.

Interestingly, for this dataset, the performance drop was not category-dependent. For instance, some of the classes, such as the mountain and kitchen, retained good classification accuracy in all scenarios whilst other classes did not. Another interesting observation was that out-door scene images were statistically the least affected by the 10.8° window occlusion. However, we observed that classification accuracy dropped more when scene images were masked with a 10.8° scotoma than when they were masked with a 10.8° window. This highlights the difference between outdoor images and natural images in terms of the location of the features.

3.2. Comparison to HMAX

We compared the performance of the En-HMAX model with that of the original HMAX model in terms of the individual class accuracies. Figure 6 shows that the En-HMAX model outperforms the HMAX model in recognising the datasets individual classes. Markers below the diagonal indicate that the En-HMAX model outperformed the HMAX model, in recognising a certain class of the dataset. We used a visual angle of 10.8° in both, scotoma and window conditions as a representation example. Of the 104 markers in Figure 6, 76 lie below the diagonal line.

3.3. Comparison to CNN

As shown in Figure 7, all tested CNNs showed similar patterns to that of the En-HMAX model (Figure 4) in the object recognition . The cross-over points in the object dataset are mainly located to the left of that in the scene dataset, suggesting that CNNs also rely more on the central image content for recognising objects. For scenes, similar prioritisation for the peripheral data was observed as the followings:

1. cross-over points of peripheral and central vision for scenes lie at the right of that for objects;
2. poor classification performance when the peripheral vision is blocked at window 13.6° for scenes.

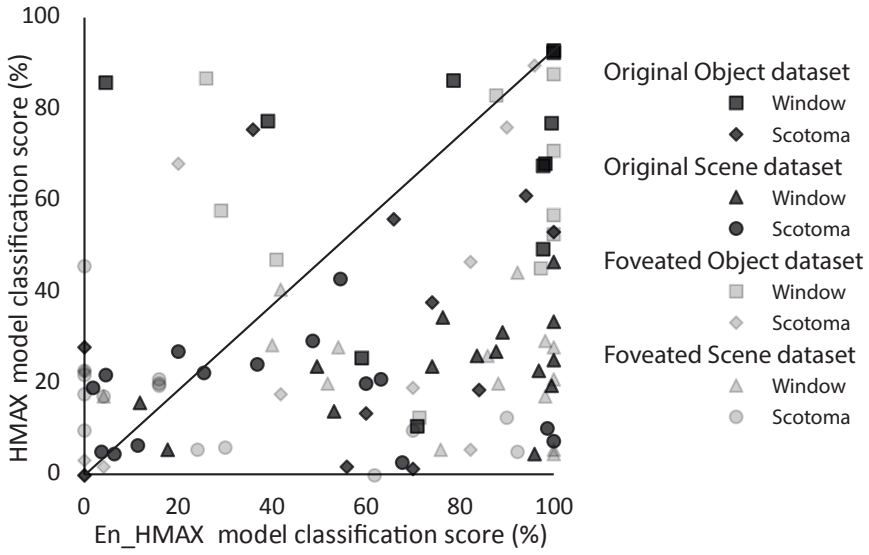

Figure 6: A comparison between the accuracy of the En-HMAX and the HMAX models. Markers of the scattered diagram represent the classes accuracies of both models. Classes below the diagonal indicate that the En-HMAX model outperforms the HMAX model. The figure shows accuracies of the 10.8° scotoma and the 10.8° window conditions. Both the original and foveated dataset were used in this analysis.

The patterns of scene recognition, however, differ from that of the En-HMAX model as the drop dramatically increased in the absence of the parafoveal vision at scotma 5°, especially for GoogLeNet and AlexNet. This suggests that the En-HMAX model utilizes the peripheral vision for recognising scenes, due to its abstract architecture. The similarity in the behavior between the CNN models and the En-HMAX model suggests that they both prioritize similar features in the images depending on its visual location. The similarity might suggest that both structures utilize sophisticated visual eccentricity biases, as the primate visual system does.

Another key difference in the behavior of the CNN models was their reduced capability in recognising foveated images, in particular, in classification of scene images with scotoma condition. For instance, when recognising scene images with 5° scotoma, the performance of the three used CNNs was significantly lower than that of the original images, as shown in Figure 7. In addition, the performance was significantly lower than that of the En-HMAX model in spite of its minimalist architecture. This observation suggests that the En-HMAX model is more robust to foveation than the CNN architecture.

3.4. Experiments 2

In Experiment 2, we tested the behavior of the En-HMAX model in classifying occluded scenes and objects with windows and scotomas of varying radii. Figure 8(A) shows that the recognition accuracy for unseen images of the scene dataset was stable to the point that more than 50% (a visual angle of 10.8°) of the image data was blocked by the scotoma. However, when the scene dataset is peripherally blocked by

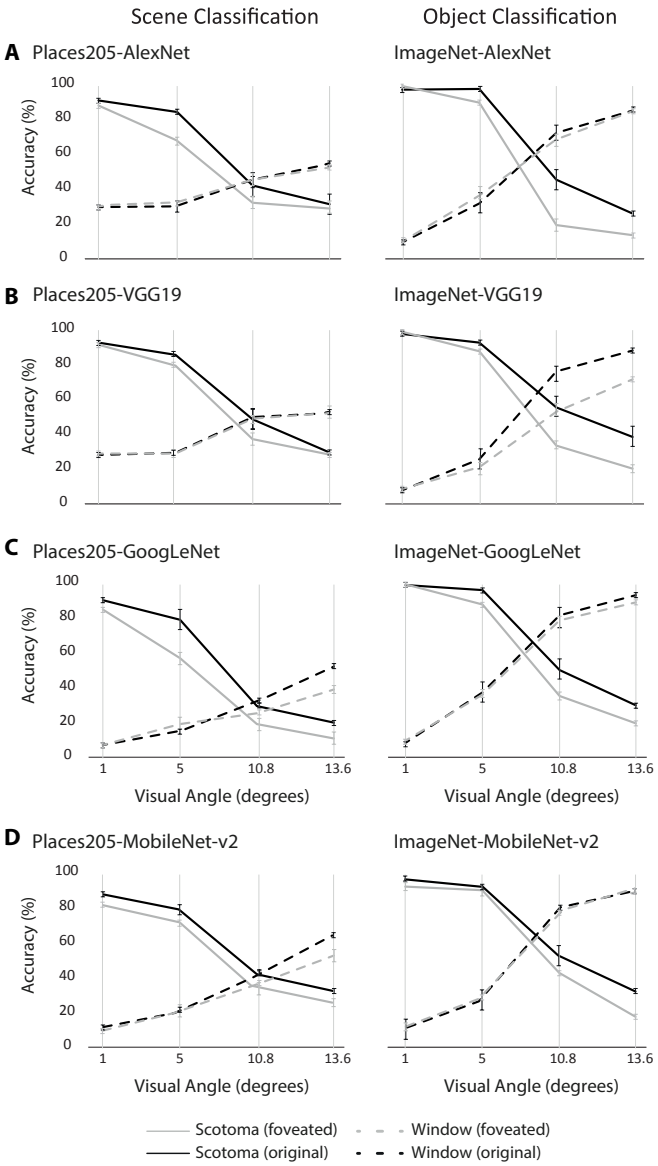

Figure 7: Replicating Experiment 1 using three well-known models of CNN, namely, AlexNet [44], VGG19 [70], GoogLeNet [72], and MobileNet-v2 [67].

the window conditions, the performance starts dropping earlier from a visual angle of 13° and downward. In Figure 8(B), the performance of object classification under the window condition is almost symmetrical. However, across the whole spectrum of visual angles, the performance under the scotoma condition was lower than that of the window in a large margin. This observation reaffirms that object recognition is more dependent on the central image content.

The performance of object classification in the window condition declined dramatically from 89% to 58% in the range of 7° to 3°. In the scene classification and in the presence of scotoma, a similar decline in

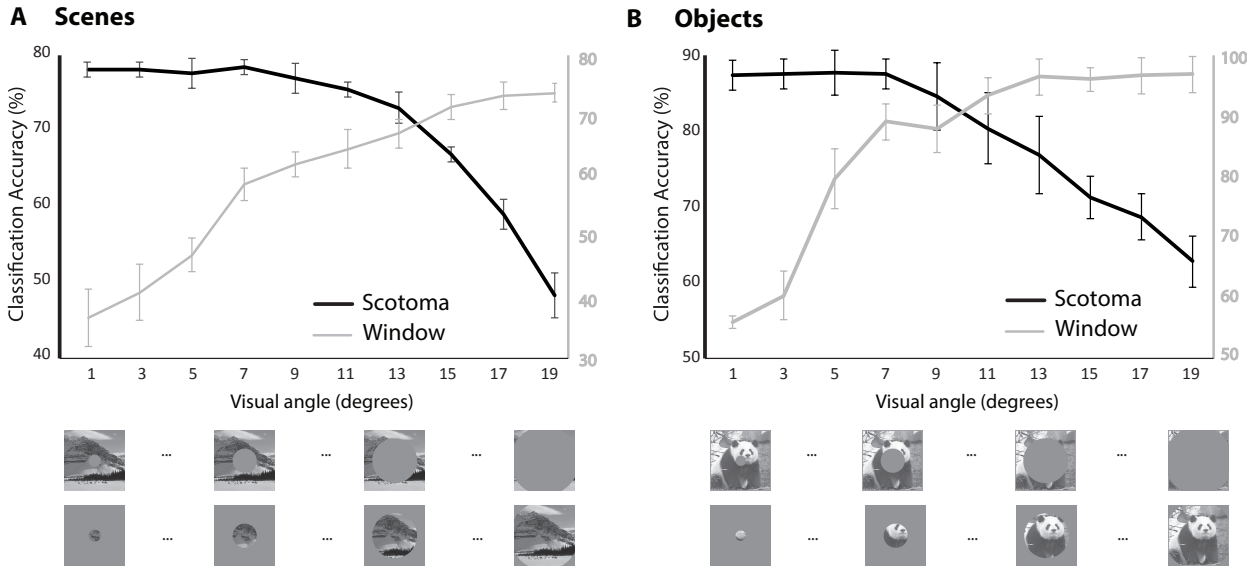

Figure 8: Classification accuracy over a range of window and scotoma visual angles. Scores are calculated for unseen scene and object images.

Table 1: Computational analysis for object recognition in Experiment 2. The table exhibits the trade-off between computation time (to process an image), average image size and the overall classification accuracy using a variety of visual angles.

Window size (degree)	**time (sec)**	**Size (kB)**	**Accuracy (%)**
19	0.057	9.8	97.23
17	0.049	9.5	97.51
15	0.044	9.3	97.12
13	0.041	8.2	93.51
11	**0.038**	**6.7**	**92.46**
9	0.034	5.2	89.45
7	0.031	3.6	90.65
5	0.029	2.4	81.91
3	0.028	1.5	59.18
1	0.028	1.1	56.94

performance took place between 13° to 17°. However, the reduction in correct classification from 73% to $\sim$ 54% was less when compared to the decline we observed for object classification. When normalized to the maximum score achieved in each condition, these reductions were 23% to 10% in the window versus scotoma, respectively. This dramatic decline in the object classification trend occurred when the visual

data around the parafoveal vision was blocked. Hence, the En-HMAX model may behave differently in this range of visual angles. With a window size of 11°, the computational requirements analysis (Tabel 1 shows that an average reduction in computation time (34.33% per-image) and image size (32.63% per-image) can be achieved by compromising only 4.99% of the total classification accuracy. For both object and scene image datasets, combining the usage of the parafoveal or peripheral areas did not significantly increase the performance.

4. Concluding Remarks

The En-HMAX model [8] is inspired by the pioneering work of Riesenhuber and Poggio [68]. It mimics the basic structure of the ventral visual stream.

Here, we investigated the contributions of peripheral versus central vision to scene and object recognition with the En-HMAX model. Also, we compared the En-HMAX behaviour with high-performing CNN structures.

Models of scene recognition [62, 74, 2, 80] show that rapid categorisation can be performed at the early perceptual stages of the visual cortex hierarchy [31, 19]. Experimental results have shown that, with a stimulus of an exposure time of 100ms, humans can categorize scenes at both the super-ordinate level (e.g. man-made versus natural) and the basic level (e.g., highway versus forest) [60].

We tested the En-HMAX model with objects and scenes datasets with varying occlusion conditions to reaffirm that peripheral image content, that is beyond 5° eccentricity, is more efficient in recognising the gist of a scene than central image content. In addition, this study showed that introducing foveation has increased the object classification performance of the En-HMAX model at 1° scotoma. However, it had no impact on recognising the gist of the scene in the absence of parafoveal vision, as showed in Figure 4.

The advantage of central vision in object recognition is mainly explained by the fact that objects are generally located in the centre of the images. This indicates that the En-HMAX model recognizes the objects within the images and not their backgrounds. Also, when normalizing performances in Experiment 2, the decline in object recognition was 13% faster than the decline in scene recognition, especially when occlusions block parafoveal $[3° - 7°]$ section of the image. This observation corroborated the importance of parafoveal vision for object recognition [56].

The prevailing advantage of the peripheral vision in scene recognition can be explained by the characteristics of scenes. The formative information of the scene is spread and compressed at the periphery of the images. Therefore, the En-HMAX model intrinsically captures the usefulness of the peripheral image content when recognising scenes. Interestingly, results suggested that outdoor man-made scene classes were less dependant on the peripheral image content. With a 10.8° window, these scene sub-types scored relatively higher performance. We speculate that the reason for this observation is that scene recognition depends on

local features within each type of scene [63]. Examples of local features are the presence of cars, pedestrians, and cyclists in a street in outdoor-man made scenes [73]. Therefore, the En-HMAX model can extract local features across man-made scene images without particularly relying on the peripheral vision. Further data and research are required to test this hypothesis.

The experiments of Wang and Cottrell [78] investigated the importance of peripheral vision on a scene image dataset. Throughout their experiments, they used complete images for training their models and occluded images for the testing. In this paper, we introduced the use of the cortex-inspired model, the En-HMAX, alongside deep learning methods. We generalized to use a scene image dataset and an object image dataset. Furthermore, we used complete and occluded images to train our models in two different experimental environments. Finally, we provided a computational analysis in terms of processing speed, image size, and total accuracy.

For completeness, we performed the classification in both parts of Experiment 1 using three established CNN architectures. In line with [78], the results showed that CNNs settings have similar preference toward peripheral image content for recognising scene images. Although foveation has proved to be efficient to alleviate the impact of adversarial examples [55], the performances were deteriorated dramatically in the presence of foveation, in particular, in the scotoma condition. The performance of the En-HMAX model was stable in the range of 1° to 5° scotoma. This difference may be explained by the biologically-informed structure of the En-HMAX setting that models rapid categorisation of the human visual cortex. Despite the simplicity of the En-HMAX architecture, it efficiently captures formative features in images. These features happen to occupy the periphery in scene images and the centre in object images. This explains how the En-HMAX model focuses on different areas when facing different image categories. Similarly, well known deep learning models, consisting of similar combinations of convolutional/pooling layers with larger scale, show a similar attitude toward object and scene images. To the best of the authors' knowledge this is the first time this observation has been made. Further human behavioral and computer modeling studies may be required to shed light on potential underlining neural or computational substrates which make the En-HMAX model robust against foveation.

In Experiment 2, we further investigated the relative importance of each region of vision for both datasets, that is, peripheral image content for scene dataset and central image content for object dataset. Blocking the less relevant image content produced the same performance pattern in both scenarios. A key outcome of this experiment may be this finding that by selectively blocking image regions, the computational requirement of image classification can be reduced which is of significant importance in real-time robotic vision applications.

The utilisation of attentional mechanism [77] in object recognition shows limited improvement to the overall accuracy in classifying unseen objects. The En-HMAX model outperforms the latter with a large margin. Other available models [13], depend on neuromorphic sensors (e.g. Event-Based cameras) to achieve visual attention. Limited research is being conducted using these sensors due to their reduced image

resolution, i.e., 128 x 128 pixels. More recent attention-based mechanisms [83, 71] are widely invested in object detection (rather than object/scene recognition). This is, to locate objects within the images by enhancing the impact of significant features and weaken background interference. The above methods do not regard the relative visual-spatial attention in recognising a scene (e.g., coast) or an object (e.g., face).

State of art models of object recognition may be too computationally expensive to run on a computer with modest specifications [51, 36, 3]. Three possibilities to overcome this problem are: 1) local processing, 2) cloud processing, and 3) a combination of the two. Cloud processing remains an important tool especially for devices with low processing capability. Most future systems may use a combination of local and cloud processing, given the increasing power of mobile graphics units and mobile connectivity. However, transferring all image data to a remote cloud may be unrealistic, due to the band-width related issues [47]. This limitation may necessitate data is either reduced or compressed locally before transmission ideally without any performance degradation. Our results showed that foveation can be an appropriate candidate for local data compression. Another important finding of our study was that the maximum classification performance, equal to when the whole image is available, can be achieved with only half of the input image content. This observation offers significant bandwidth saving and data reduction and can be an important factor to solve the band-width dilemmas in real-time Cloud-based object recognition applications [47].

Our current research includes the development of a hierarchical system for object recognition to use in robotic vision applications, for instance, in vision enabled prosthetics [29, 28]. The prior knowledge of the task (e.g., scene understanding or recognizing objects in an office) may result in leveraging the relevant area of vision [6]. For each task, the hierarchy can start with a scene recognition stage. Determining the type of scene, e.g. indoor or outdoor, will give the prosthesis an insight into the nature of objects within that scene. At a second layer, objects and scenes may be classified with specialized classifiers.

Appendix

The object classes in this dataset and the number of samples in each class $(\cdot)$ are: cars (1300), dollar-bills (900), Faces (1280), Garfields (910), skates (999), motorbikes (1280), pagodas (1001), pandas (1020), scissors (1250), trilobites (952), chairs (1222), where the number in the parentheses is the number of images in the class. The scene image classes are: bedrooms (1223), suburbs (1373), industrials (1184), kitchens (1294), livingrooms (1216), coasts (1199), forests (1260), highways (1273), cities (1207), mountains (1273), Country-sides (1248), streets (1227), buildings (1288), offices (1237), stores (1283). The CNN models were pre-trained with 1,281,167 object images and 2,500,000 scene images depending on the task.

Conflicts of interest

There is no conflict to interest.

Funding

The work of A. Alameer is supported by Newcastle University. The work of K. Nazarpour is supported by the Engineering and Physical Sciences Research Council, U.K., grants EP/M025977/1 and EP/M025594/1.

References

[1] Abolghasemi, V., Chen, M., Alameer, A., Ferdowsi, S., Chambers, J., and Nazarpour, K. Incoherent dictionary pair learning: Application to a novel open-source database of chinese numbers. *IEEE Signal Processing Letters 25*, 4 (2018), 472–476.

[2] Aditya, S., Yang, Y., Baral, C., Aloimonos, Y., and Fermüller, C. Image understanding using vision and reasoning through scene description graph. *Computer Vision and Image Understanding* (2017).

[3] Alameer, A., and Akkar, H. Ecg signal diagnoses using intelligent systems based on fpga. *Engineering and Technology Journal 31*, 7 Part (A) Engineering (2013), 1351–1364.

[4] Alameer, A., Degenaar, P., and Nazarpour, K. Biologically-inspired object recognition system for recognizing natural scene categories. In *2016 International Conference for Students on Applied Engineering (ICSAE)* (2016), IEEE, pp. 129–132.

[5] Alameer, A., Degenaar, P., and Nazarpour, K. Processing occlusions using elastic-net hierarchical max model of the visual cortex. In *2017 IEEE International Conference on INnovations in Intelligent SysTems and Applications (INISTA)* (2017), IEEE, pp. 163–167.

[6] Alameer, A., Degenaar, P., and Nazarpour, K. Context-based object recognition: Indoor versus outdoor environments. In *Science and Information Conference* (2019), Springer, pp. 473–490.

[7] Alameer, A., Ghazaei, G., Degenaar, P., Chambers, J. A., and Nazarpour, K. Object recognition with an elastic net-regularized hierarchical MAX model of the visual cortex. *IEEE Sig. Process. Lett. 23*, 8 (2016), 1062–1066.

[8] Alameer, A., Ghazaei, G., Degenaar, P., and Nazarpour, K. An elastic net-regularized HMAX model of visual processing. In *Proc. 2^{nd} IET Int. Conf. Intelligent Sig. Process.* (2015), pp. 1–4.

[9] Alameer, A. M. A. *Biologically-inspired hierarchical architectures for object recognition*. PhD thesis, Newcastle University, 2018.

[10] Arbib, M. A., and Bonaiuto, J. J. *From Neuron to Cognition Via Computational Neuroscience*. MIT Press, 2016.

[11] Baldassano, C., Fei-Fei, L., and Beck, D. M. Pinpointing the peripheral bias in neural scene-processing networks during natural viewing. *J. Vis. 16*, 2 (2016).

[12] Bolduc, M., and Levine, M. D. A real-time foveated sensor with overlapping receptive fields. *Real-Time Imaging 3*, 3 (1997), 195–212.

[13] Cannici, M., Ciccone, M., Romanoni, A., and Matteucci, M. Attention mechanisms for object recognition with event-based cameras. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)* (2019), pp. 1127–1136.

[14] Carrasco, M., McElree, B., Denisova, K., and Giordano, A. M. Speed of visual processing increases with eccentricity. *Nat. Neurosci. 6*, 7 (2003), 699.

[15] Chang, Y. Research on de-motion blur image processing based on deep learning. *Journal of Visual Communication and Image Representation* (2019).

[16] Cheng, G., Han, J., Zhou, P., and Xu, D. Learning rotation-invariant and fisher discriminative convolutional neural networks for object detection. *IEEE Transactions on Image Processing 28*, 1 (2018), 265–278.

[17] Cheng, G., Yang, C., Yao, X., Guo, L., and Han, J. When deep learning meets metric learning: Remote sensing image scene classification via learning discriminative cnns. *IEEE transactions on geoscience and remote sensing 56*, 5 (2018), 2811–2821.

[18] Cheng, G., Zhou, P., and Han, J. Learning rotation-invariant convolutional neural networks for object detection in vhr optical remote sensing images. *IEEE Transactions on Geoscience and Remote Sensing 54*, 12 (2016), 7405–7415.

[19] Coggan, D. D., Allen, L. A., Farrar, O. R., Gouws, A. D., Morland, A. B., Baker, D. H., and Andrews, T. J. Differences in selectivity to natural images in early visual areas (V1–V3). *Sci. Rep. 7* (2017), 2444.

[20] Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., and Fei-Fei, L. ImageNet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on* (2009), pp. 248–255.

[21] Epstein, R., Harris, A., Stanley, D., and Kanwisher, N. The Parahippocampal place area: Recognition, navigation, or encoding? *Neuron 23*, 1 (1999), 115–125.

[22] Farzmahdi, A., Rajaei, K., Ghodrati, M., Ebrahimpour, R., and Khaligh-Razavi, S.-M. A specialized face-processing model inspired by the organization of monkey face patches explains several face-specific phenomena observed in humans. *Sci. Rep. 6* (2016), 25025.

[23] Fei-Fei, L., Fergus, R., and Perona, P. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comp. Vis. Imag. Underst. 106*, 1 (2007), 59–70.

[24] Fukushima, K., and Miyake, S. Neocognitron: A new algorithm for pattern recognition tolerant of deformations and shifts in position. *Patt. Recog. 15*, 6 (1982), 455–469.

[25] Geisler, W. S., and Perry, J. S. Variable-resolution displays for visual communication and simulation. *SID Symposium Digest of Technical Papers 30*, 1 (1999), 420–423.

[26] Geisler, W. S., and Perry, J. S. Real-time simulation of arbitrary visual fields. In *Proc. Symp. Eye Tracking Research and Applications* (2002), pp. 83–87.

[27] Geisler, W. S., Perry, J. S., and Najemnik, J. Visual search: The role of peripheral information measured using gaze-contingent displays. *J. Vis. 6*, 9 (2006), 1–1.

[28] Ghazaei, G., Alameer, A., Degenaar, P., Morgan, G., and Nazarpour, K. An exploratory study on the use of convolutional neural networks for object grasp classification.

[29] Ghazaei, G., Alameer, A., Degenaar, P., Morgan, G., and Nazarpour, K. Deep learning-based artificial vision for grasp classification in myoelectric hands. *J. Neural Eng. 14*, 3 (2017), 23–36.

[30] Gomez, J., Pestilli, F., Witthoft, N., Golarai, G., Liberman, A., Poltoratski, S., Yoon, J., and Grill-Spector, K. Functionally defined white matter reveals segregated pathways in human ventral temporal cortex associated with category-specific processing. *Neuron 85*, 1 (2015), 216–227.

[31] Greene, M. R., and Oliva, A. The briefest of glances the time course of natural scene understanding. *Psychol. Sci. 20*, 4 (2009), 464–472.

[32] Grill-Spector, K., and Malach, R. The human visual cortex. *Annual Rev. Neurosci. 27* (2004), 649–677.

[33] Gulcehre, C., Cho, K., Pascanu, R., and Bengio, Y. Learned-norm pooling for deep feedforward and recurrent neural networks. In *Joint European Conf. Mach. Learn. Knowl. Disc. in Databases* (2014), pp. 530–546.

[34] Henderson, J. M., and Hollingworth, A. The role of fixation position in detecting scene changes across saccades. *Psychol. Sci. 10*, 5 (1999), 438–443.

[35] Henderson, J. M., McClure, K. K., Pierce, S., and Schrock, G. Object identification without foveal vision: Evidence from an artificial scotoma paradigm. *Atten., Percep., and Psychoph. 59*, 3 (1997), 323–346.

[36] Holliday, A., Barekatain, M., Laurmaa, J., Kandaswamy, C., and Prendinger, H. Speedup of deep learning ensembles for semantic segmentation using a model compression technique. *Computer Vision and Image Understanding 164* (2017), 16–26.

[37] Hubel, D. H., and Wiesel, T. N. Receptive fields of single neurones in the cat's striate cortex. *J. Physiol. 148*, 3 (1959), 574–591.

[38] Kaiser, P. K. *The joy of visual perception.* York University, 2009.

[39] Kanwisher, N., McDermott, J., and Chun, M. M. The fusiform face area: A module in human extrastriate cortex specialized for face perception. *J. Neurosci. 17*, 11 (1997), 4302–4311.

[40] Karklin, Y., and Lewicki, M. S. A hierarchical bayesian model for learning nonlinear statistical regularities in nonstationary natural signals. *Neural Comp. 17*, 2 (2005), 397–423.

[41] Kevin O'Regan, J., Deubel, H., Clark, J. J., and Rensink, R. A. Picture changes during blinks: Looking without seeing and seeing without looking. *Vis. Cognition 7*, 1-3 (2000), 191–211.

[42] Kheradpisheh, S. R., Ghodrati, M., Ganjtabesh, M., and Masquelier, T. Deep networks can resemble human feed-forward vision in invariant object recognition. *Sci. Rep. 6* (2016), 32672.

[43] Kingma, D. P., and Ba, J. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* (2014).

[44] Krizhevsky, A., Sutskever, I., and Hinton, G. E. ImageNet classification with deep convolutional neural networks. In *Adv. in NIPS*. 2012, pp. 1097–1105.

[45] Larson, A. M., and Loschky, L. C. The contributions of central versus peripheral vision to scene gist recognition. *J. Vis. 9*, 10 (2009), 46–53.

[46] Lazebnik, S., Schmid, C., and Ponce, J. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Proc. IEEE Conf. Comp. Vis. Patt. Recog.* (2006), pp. 2169–2178.

[47] Lee, J., Wang, J., Crandall, D., Šabanović, S., and Fox, G. Real-time, cloud-based object detection for unmanned aerial vehicles. In *Proc. IEEE Int. Conf. Robotic Computing* (2017), pp. 36–43.

[48] Li, F.-F., and Perona, P. A bayesian hierarchical model for learning natural scene categories. In *Proc. IEEE Conf. Comp. Vis. Patt. Recog.* (2005), pp. 524–531.

[49] Li, Y., Wu, W., Zhang, B., and Li, F. Enhanced hmax model with feedforward feature learning for multiclass categorization. *Frontiers in computational neuroscience 9* (2015), 123.

[50] Loschky, L. C., Sethi, A., Simons, D. J., Pydimarri, T. N., Ochs, D., and Corbeille, J. L. The importance of information localization in scene gist recognition. *J. Exp. Psychology: Human Percept. Perform. 33*, 6 (2007).

[51] Lu, K., Chen, J., Little, J. J., and He, H. Lightweight convolutional neural networks for player detection and classification. *Computer Vision and Image Understanding* (2018).

[52] Lu, X., Ji, W., Li, X., and Zheng, X. Bidirectional adaptive feature fusion for remote sensing scene classification. *Neurocomputing 328* (2019), 135–146.

[53] Lu, X., Li, X., and Mou, L. Semi-supervised multitask learning for scene recognition. *IEEE transactions on cybernetics 45*, 9 (2014), 1967–1976.

[54] Lu, X., Zheng, X., and Yuan, Y. Remote sensing scene classification by unsupervised representation learning. *IEEE Transactions on Geoscience and Remote Sensing 55*, 9 (2017), 5148–5157.

[55] Luo, Y., Boix, X., Roig, G., Poggio, T., and Zhao, Q. Foveation-based mechanisms alleviate adversarial examples. *arXiv preprint arXiv:1511.06292* (2015).

[56] Malach, R., Levy, I., and Hasson, U. The topography of high-order human object areas. *Tr. Cog. Sci. 6*, 4 (2002), 176–184.

[57] Massé, B., Ba, S., and Horaud, R. Tracking gaze and visual focus of attention of people involved in social interaction. *IEEE Trans. Patt. Anal. Mach. Intell. PP*, 99 (2017), 1–1.

[58] McCandliss, B. D., Cohen, L., and Dehaene, S. The visual word form area: expertise for reading in the fusiform gyrus. *Tre. Cog. Sci. 7*, 7 (2003), 293–299.

[59] McConkie, G. W., and Rayner, K. The span of the effective stimulus during a fixation in reading. *Atten., Percep., and Psychoph. 17*, 6 (1975), 578–586.

[60] Miguel Thibaut, Thi Ha Tran, S. S., and Boucart, M. The contribution of central and peripheral vision in scene categorization: A study on people with central vision loss. *Vis. Res. 98* (2014), 46–53.

[61] Mormann, F., Kornblith, S., Cerf, M., Ison, M. J., Kraskov, A., Tran, M., Knieling, S., Quiroga, R. Q., Koch, C., and Fried, I. Scene-selective coding by single neurons in the human Parahippocampal cortex. *Proc. Nat. Acad. Sci.* (2017), 201608159.

[62] Oliva, A., and Torralba, A. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comp. Vis. 42*, 3 (2001), 145–175.

[63] Oliva, A., and Torralba, A. Building the gist of a scene: The role of global image features in recognition. *Prog. Brain Res. 155* (2006), 23–36.

[64] Olshausen, B. A., and Field, D. J. Emergence of simple-cell receptive field properties by learning a sparse code for natural images. *Nature 381*, 6583 (1996), 607.

[65] Reingold, E. M., Loschky, L. C., McConkie, G. W., and Stampe, D. M. Gaze-contingent multiresolutional displays: An integrative review. *Human Factors 45*, 2 (2003), 307–328.

[66] Russakovsky, O., et al. ImageNet large scale visual recognition challenge. *Int. J. Comp. Vis. 115*, 3 (2015), 211–252.

[67] Sandler, M., Howard, A., Zhu, M., Zhmoginov, A., and Chen, L.-C. Mobilenetv2: Inverted residuals and linear bottlenecks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (2018), pp. 4510–4520.

[68] Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., and Poggio, T. Robust object recognition with cortex-like mechanisms. *IEEE Trans. Patt. Anal. Mach. Intell. 29*, 3 (2007), 411–426.

[69] Shipley, T. F., and Kellman, P. J. *From fragments to objects: Segmentation and grouping in vision.* North Holland, 2001.

[70] Simonyan, K., and Zisserman, A. Very deep convolutional networks for large-scale image recognition. *CoRR abs/1409.1556* (2014).

[71] Symeonidis, G., and Tefas, A. Recurrent attention for deep neural object detection. In *Proceedings of the 10th Hellenic Conference on Artificial Intelligence* (2018), p. 3.

[72] Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., and Rabinovich, A. Going deeper with convolutions. In *Proc. IEEE Conf. Comp. Vis. Patt. Recog.* (2015), pp. 1–9.

[73] Teichman, A., and Thrun, S. Practical object recognition in autonomous driving and beyond. In *Adv. Robotics and its Social Impacts (ARSO), 2011 IEEE Workshop on* (2011), pp. 35–38.

[74] Torralba, A., and Oliva, A. Statistics of natural image categories. *Network: Computation in Neural Systems 14* (2003), 391–412.

[75] Ullman, S., Assif, L., Fetaya, E., and Harari, D. Atoms of recognition in human and computer vision. *Proc. Nat. Acad. Sci. 113*, 10 (2016), 2744–2749.

[76] Van Diepen, P. M., and Wampers, M. Scene exploration with fourier-filtered peripheral information. *Perception 27*, 10 (1998), 1141–1151.

[77] Walther, D., Itti, L., Riesenhuber, M., Poggio, T., and Koch, C. Attentional selection for object recognition—a gentle way. In *International workshop on biologically motivated computer vision* (2002), Springer, pp. 472–479.

[78] Wang, P., and Cottrell, G. W. Central and peripheral vision for scene recognition: a neurocomputational modeling exploration. *Jour. of Vis. 17*, 4 (2017), 9–9.

[79] Yang, J., Yu, K., Gong, Y., and Huang, T. Linear spatial pyramid matching using sparse coding for image classification. In *Proc. IEEE Conf. Comp. Vis. Patt. Recog.* (2009), pp. 1794–1801.

[80] Yin, W., Xu, D., Wang, Z., Zhao, Z., Chen, C., and Yao, Y. Perceptually learning multi-view sparse representation for scene categorization. *Journal of Visual Communication and Image Representation* (2019).

[81] Young, R. A. The Gaussian derivative model for spatial vision: I. retinal mechanisms. *Spatial Vis. 2*, 4 (1987), 273–293.

[82] Zalluhoglu, C., and Ikizler-Cinbis, N. Region based multi-stream convolutional neural networks for collective activity recognition. *Journal of Visual Communication and Image Representation* (2019).

[83] Zhang, Y., Chen, Y., Huang, C., and Gao, M. Object detection network based on feature fusion and attention mechanism. *Future Internet 11*, 1 (2019), 9.

[84] Zhou, B., Lapedriza, A., Xiao, J., Torralba, A., and Oliva, A. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems* (2014), pp. 487–495.

[85] Zou, H., and Hastie, T. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society: Series B (Statistical Methodology) 67*, 2 (2005), 301–320.